

Jumping into buckets, or How to decontaminate overlapping fat jets

Koichi Hamaguchi,^{1,2} Seng Pei Liew,¹ and Martin Stoll¹

¹*Department of Physics, University of Tokyo,
Bunkyo-ku, Tokyo 113-0033, Japan.*

²*Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU),
University of Tokyo, Kashiwa 277-8583, Japan*

Abstract

At the LHC, tagging boosted heavy particle resonances which decay hadronically, such as top quarks and Higgs bosons, can play an essential role in new physics searches. In events with high multiplicity, however, the standard approach to tag boosted resonances by a large-radius fat jet becomes difficult because the resonances are not well-separated from other hard radiation. In this paper, we propose a different approach to tag and reconstruct boosted resonances by using the recently proposed mass-jump jet algorithm. A key feature of the algorithm is the flexible radius of the jets, which results from a terminating veto that prevents the recombination of two hard prongs if their combined jet mass is substantially larger than the masses of the separate prongs. The idea of collecting jets in “buckets” is also used. As an example, we consider the fully hadronic final state of pair-produced vectorlike top partners at the LHC, $pp \rightarrow T\bar{T} \rightarrow t\bar{t}HH$, and show that the new approach works well. We also show that tagging and kinematic reconstruction of boosted top quarks and Higgs bosons are possible with good quality even in these very busy final states. The vectorlike top partners are kinematically reconstructed, which allows their direct mass measurement.

I. INTRODUCTION

The high-energy frontier of particle physics is often probed at hadronic colliders, such as Tevatron and the Large Hadron Collider (LHC). Most of the time, collisions at hadron colliders produce coloured partons that undergo showering and hadronization resulting in a large number of final-state hadrons. In order to extract momentum, energy, and other information of the hard-scattering partons, it is necessary to cluster the hadrons observed at the detectors into jets. Studying jets and their clustering procedure (jet algorithm) are therefore the keys to understanding the physics at hadron colliders.

It is particularly interesting to consider the clustering of hadrons originating from boosted heavy particle resonances such as top quarks and Higgs bosons, given that the LHC restarts with an increased energy $\sqrt{s} = 13\text{-}14\text{ TeV}$ and will produce such resonances copiously. The construction of large-radius “fat” jets has become a common approach to deal with such scenarios, facilitated by progress in tagging of boosted resonances (see e.g. Refs. [1, 2] and references therein). By investigating the substructure of a fat jet, more information on the energy deposit pattern is available compared to separately resolved small-radius jets. Furthermore, the classical problem of finding an optimal jet radius [3–5] is avoided and jet combinatorics are significantly reduced.

A more challenging situation arises when multiple heavy resonances are produced simultaneously. Such processes lead to very busy final states where the heavy particles under consideration (such as top and Higgs) as well as their daughter particles are not well-separated. (See Fig. 1.) As a result, fat jets merge in most events, and a majority of the jets contain decay products from more than one resonance. Such a scenario is not adequately addressed by most tagging algorithms based on (isolated) fat jets.

In the present work, we suggest a new framework of jet tagging that allows particle reconstruction with good quality compared to traditional methods in such a busy hadronic environment. A key ingredient is the “mass-jump” jet clustering algorithm [6], which is an extension of the “mass-drop” subjet identification in fat jets as employed in the HEPTop-Tagger [7]. It has been shown that the mass-jump algorithm gives competitive performance, but now the “sub”jets are formed directly without the definition of an intermediate fat jet. Mass-jump clustering harnesses the advantage of fat jet substructure algorithms of resolving small jets without reference to a fixed radius.

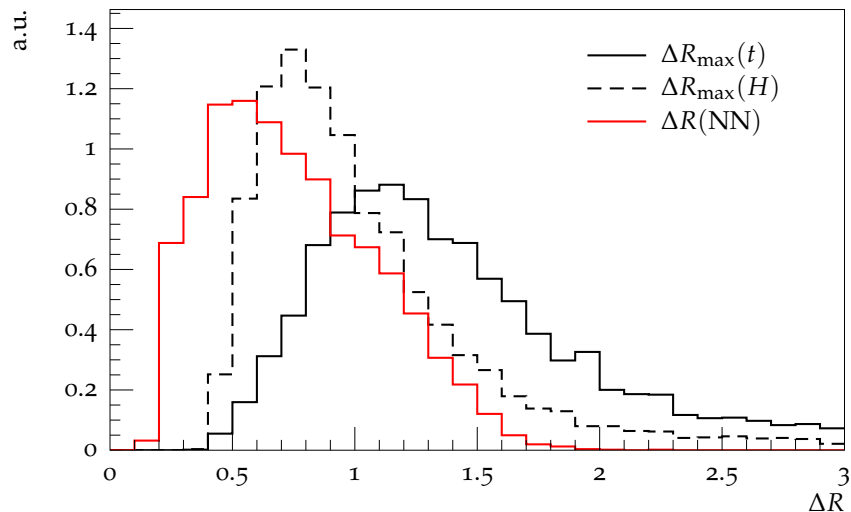


FIG. 1: Angular distances between daughter particles from top and Higgs decays, in the benchmark process $pp \rightarrow T\bar{T} \rightarrow t\bar{t}HH \rightarrow 10$ jets at the LHC with $\sqrt{s} = 14$ TeV, with vectorlike top mass $m_T = 1$ TeV (parton level, arbitrary units). The largest R -distance between the daughters of the same top quark (Higgs boson) is denoted by $\Delta R_{\max}(t)$ ($\Delta R_{\max}(H)$) and plotted with black solid (dashed) lines. The minimal R -distance between the nearest neighbour daughters coming from different mothers $\Delta R(\text{NN})$ is depicted by the red line. See Sec. III for details.

The absence of the fat jet, however, reintroduces the issue of large combinatorics in such busy environments, since the decay products of the heavy resonances cannot be disentangled a priori. To facilitate event reconstruction, the idea of collecting jets into separate “buckets” [8, 9] is applied, which allows efficient assignment of jets to their respective resonances.

The jet-tagging method proposed here is applicable to a broad range of Standard Model (SM) and beyond the SM phenomena at hadron colliders. There are indeed important SM processes which involve decays of multiple heavy particles, resulting in a busy hadronic environment. A prime example is the associated production of a Higgs boson with two top quarks ($pp \rightarrow t\bar{t}H$), which has attracted attention as this channel opens up the opportunity to measure directly the Higgs-top Yukawa coupling, an essential probe towards understanding the Higgs sector.

Some models of supersymmetry (SUSY) also predict large multiplicity of jets with little or no missing transverse energy (MET). For example, assuming that the gluino is the lightest

SUSY particle, it can decay into a top quark and jets when baryonic R-parity associated to the third generation quark is violated [10]. This leads to a multijet final state when the top quark decays hadronically ($\tilde{g}\tilde{g} \rightarrow ttjjjj$). Another example is the stealth SUSY, where the top and the lighter stop (\tilde{t}_1) masses are almost degenerate, leading to final states without significant MET [11]. The heavier stop (\tilde{t}_2) has a model-dependent decay branching ratio to the Z or Higgs boson ($\tilde{t}_2 \rightarrow \tilde{t}_1 Z/H$). The hadronic mode of this decay again leads to multijet final states with little MET ($\tilde{t}_2\tilde{t}_2^* \rightarrow t\bar{t}(H/Z)(H/Z)$). Our method may allow to fully reconstruct the underlying new particles in such models, too.

In order to illustrate the strength of our jet tagging method, we investigate a simplified version of heavy particle production topology in this paper, i.e. we consider the fully hadronic final state of pair-produced vectorlike top partners at the LHC ($pp \rightarrow T\bar{T} \rightarrow t\bar{t}HH$). In particular, we study the performance of our taggers at the 14 TeV LHC with a vectorlike top of mass around 1 TeV as our benchmark scenario. Studies of fully hadronic final states in similar processes have been based on fat jet substructure [12–15], including experimental searches at 8 TeV by CMS [16]. Current exclusion bounds on the vectorlike top mass are $m_T \gtrsim 700 - 950$ from ATLAS [17–19] and $m_T \gtrsim 690 - 910$ from CMS [20–22], depending on the assumed branching fractions.

This paper is arranged as follows: In Section II, we review the essential tools used in our analysis: the mass-jump clustering and the bucket algorithms. We then apply our method to the benchmark scenario of the fully hadronic decay of pair-produced vectorlike tops in Section III. The performance of the involved top and Higgs taggers is investigated in Section IV. We conclude our findings in Section V.

II. RECAP: MASS-JUMP JET CLUSTERING AND BUCKETS

In this paper, we investigate a new approach of analyzing high-multiplicity final states based on separately resolved jets. We try to answer the two key questions that arise in such an analysis: (1) which algorithm to use to construct the jets, and (2) how to reduce the sheer combinatorial choices of assigning the jets to the resonance particles of the process. We examine the first question by comparing the mass-jump algorithm [6] with the corresponding jet clustering algorithm of the generalized k_T family. The latter question is addressed by the bucket algorithm introduced in Refs. [8, 9]. Both recent techniques are briefly reviewed

in the remainder of this section.

A. Jet clustering with a mass-jump veto

In the commonly used generalized- k_T family of jet clustering algorithms, jets are constructed by sequential recombination of input particles until a certain angular distance is reached, the jet radius R . As a result, all jet centres are mutually separated by $\Delta R(j_1, j_2) = \sqrt{(\phi_1 - \phi_2)^2 + (y_1 - y_2)^2} > R$ and the angular spread of each jet is roughly $\leq R$. ϕ and y are the azimuthal angle and rapidity, respectively. The choice of a large parameter R can lead to radiation from two (or more) hard partons ending up in the same jet (splash-in). On the other hand, if the radius is too small, not all final-state radiation of a hard parton will be captured by the jet (splash-out). In both cases, jet-parton correspondence is disturbed.

The mass-jump clustering algorithm [6] addresses this problem by implementing a flexible jet radius based on an intrinsic jet property (jet mass) as well as the topology of jets in its vicinity. Two parameters are introduced accordingly, the jet mass threshold μ and the mass-ratio parameter θ . Jet clustering starts from a set of input particles, which are labelled *active* protojets. A distance metric for protojets j_i is defined as

$$d_{j_1 j_2} = \frac{\Delta R_{j_1 j_2}^2}{R^2} \min [p_{j_1 \perp}^{2n}, p_{j_2 \perp}^{2n}] , \quad d_{j_1 B} = p_{j_1 \perp}^{2n} , \quad (1)$$

where $n = 1$ corresponds to the k_T algorithm [23–25], $n = 0$ to the Cambridge/Aachen algorithm [26, 27], and $n = -1$ to the anti- k_T algorithm [28]. The sequential recombination algorithm then proceeds as follows [6]:

1. Find the smallest $d_{j_a j_b}$ among active protojets, including $d_{j_a B}$; if it is given by a beam distance, $d_{j_a B}$, label j_a *passive* and repeat step 1.
2. Combine j_a and j_b by summing their four-momenta, $p_{j_a+j_b} = p_{j_a} + p_{j_b}$ (E -scheme). If the new jet is still light, $m_{j_a+j_b} < \mu$, replace j_a and j_b by their combination in the set of active protojets and go back to step 1.

Otherwise check the mass-jump criterion: If $\theta \cdot m_{j_a+j_b} > \max[m_{j_a}, m_{j_b}]$ label j_a and j_b *passive* and go back to step 1.

3. Mass jumps can also appear between an active and a passive protojet. To examine this
 - a. Find the passive protojet j_n which is closest to j_a in terms of the metric d and is not isolated, $d_{j_a j_n} < d_{j_n B}$.
 - b. Then check if these two protojets would have been recombined if j_n had not been rendered passive by a previous veto, i.e. $d_{j_a j_n} < d_{j_a j_b}$.
 - c. Finally check the mass-jump criterion, $m_{j_a+j_n} \geq \mu$ and $\theta \cdot m_{j_a+j_n} > \max[m_{j_a}, m_{j_n}]$.

If all these criteria for the veto are fulfilled, label j_a *passive*. Do the same for j_b . If either of j_a or j_b turned passive, go back to step 1.

4. No mass-jump has been found, so replace j_a and j_b by their combination in the set of active protojets. Go back to step 1.

Clustering terminates when there are no more active protojets left. Passive protojets are then labelled jets. Note that for $\theta = 0$ or $\mu = \infty$ standard sequential clustering without veto is recovered, which is reduced to steps 1 and 4. Jet clustering can be k_T -like, C/A-like, or anti- k_T -like, depending on the metric chosen [see Eq. (1)].

In Ref. [6] it has been shown that the mass-jump clustering algorithm can be useful to resolve the close-by subjects of boosted top quarks. At the same time, isolated jets are hardly affected by the veto if μ and θ are not chosen too aggressively. These properties qualify the mass-jump algorithm as a suitable candidate for processes with very busy final states where this flexibility is essential.

The mass-jump algorithm described above is the first member of the family of jet clustering algorithms with a terminating veto and has been made publicly available as part of the FASTJET contribution package [29]. The plugin is dubbed `ClusteringVetoPlugin` and accepts any user-defined veto function. Its exemplary usage is illustrated within the package.

B. The bucket algorithm

In high-multiplicity events, the assignment of jets to their respective heavy resonances can easily get out of hand. There are $6!/(3!3!) = 20$ possible combinations to assign six jets

to two top quarks, and for eight jets coming from two tops and a Higgs boson this number already reaches $8!/(3!3!2!) = 560$. The bucket algorithm [8, 9] was proposed in the context of these two final states and introduces a “bucket” for each top quark (and one additional bucket of unassigned jets B_{ISR}), into which the jets are allocated. For a bucket B_i the metric

$$\Delta_{B_i} = |m_{B_i} - m_t| \quad \text{with} \quad m_{B_i}^2 = \left(\sum_{j \in B_i} p_j \right)^2 \quad (2)$$

measures the similarity of a collection of jets inside the bucket with a top quark. In Refs. [8, 9], the combination is determined by minimizing a global χ^2 -like metric defined as

$$\Delta^2 = \omega \Delta_{B_1}^2 + \Delta_{B_2}^2, \quad (3)$$

and choosing a large $\omega = 100$ effectively decouples the two buckets. Thereby $\Delta_{B_1} < \Delta_{B_2}$ holds and the problem of unfeasible combinatorics is circumvented because the buckets can be filled independently. Top tagging is performed by imposing cuts on each bucket later on. Whereas in the original proposal the number of jets inside each bucket is not fixed, in this paper we require strictly three jets in each top bucket, and also introduce Higgs buckets that contain exactly two jets.¹

We apply the bucket algorithm in conjunction with mass-jump jet clustering as well as conventional jet clustering for comparison. Our benchmark scenario is given by $t\bar{t}HH$ production from a pair of vectorlike tops. Clearly the naive combinatorics are overwhelming even for the minimal final-state multiplicity of ten jets: $10!/(3!3!2!2!) = 25200$. To tackle this problem we formally define a global metric

$$\Delta^2 = \omega_1 \Delta_{B_{t1}}^2 + \omega_2 \Delta_{B_{t2}}^2 + \omega_3 \Delta_{B_{H1}}^2 + \omega_4 \Delta_{B_{H2}}^2 \quad (4)$$

and explicitly decouple the four buckets by choosing the (positive) weights such that

$$\frac{\omega_{i+1}}{\omega_i} = +0 \quad i = 1..3. \quad (5)$$

Therefore, the buckets are filled separately in order $(B_{t1}, B_{t2}, B_{H1}, B_{H2})$ and the computational load is reduced to only $10!/7!/3! + 7!/4!/3! + 4!/2!/2! = 161$ comparisons.² In reality

¹ In the top quark rest frame, in a large fraction of events, one of the decay products from $t \rightarrow bW^+ \rightarrow bj\bar{j}$ carries low transverse momentum and thus fails to be reconstructed as a jet. As this paper is concerned with boosted top quarks from a heavy resonance decay, this problem does not occur.

² If the Higgs buckets are filled before the top buckets, the combination is further reduced to $10!/8!/2! + 8!/6!/2! + 6!/3!/3! = 93$. However, it would increase the wrong assignments for both the signal and background. See also discussion in Section IV B.

the number of jets will often be larger than the minimum of ten, where the speedup indicated here becomes even more prominent. A detailed description of our specific algorithm is given in the next section. We address possible issues related to the explicit decoupling of the buckets in Section IV B.

III. BENCHMARK SCENARIO: TEN-JET FINAL STATE FROM VECTORLIKE TOP PAIR PRODUCTION

A. Model and event generation

In this section, we illustrate our approach by using a simple model with a vectorlike top partner. We extend the SM Lagrangian by adding a vectorlike top T that interacts with the SM top t and Higgs H ,³

$$\mathcal{L} = \mathcal{L}_{\text{SM}} + \bar{T}(i\not{D} - m_T)T + y_T H \bar{t} T + h.c. . \quad (6)$$

We assume that the vectorlike top decays exclusively to a top and a Higgs. The mass of the vectorlike top in consideration is $m_T = 0.8 - 1.2$ TeV. The mass of the top is taken to be 173 GeV. The SM Higgs has mass 126 GeV and decays to $b\bar{b}$ with branching ratio 56 %.

MADGRAPH5_AMC@NLO 2.2.1 [30] is used for generating parton-level events, which undergo hadronization and showering through PYTHIA 6.426 [31]. DELPHES 3.1.2 [32] with parameters tuned to the ATLAS detector is used for fast detector simulation.

The relevant SM background processes for our analysis (cf. Sec. III B) and their respective NLO K-factors are $pp \rightarrow t\bar{t}$ (1.61 [33]), $pp \rightarrow t\bar{t}b\bar{b}$ (1.77 [34]), $pp \rightarrow t\bar{t}H$ (1.10 [35]), and $pp \rightarrow b\bar{b}b\bar{b}$ (1.40 [36]). All final-state top quarks and Higgs bosons are decayed hadronically within MADGRAPH5_AMC@NLO. The following generator-level cuts are imposed: minimum transverse momentum of each outgoing parton $p_{\perp} \geq 20$ GeV, angular separation between outgoing light quarks and between a light quark and a bottom quark $\Delta R_{jj}, \Delta R_{jb} \geq 0.2$, and angular separation between a pair of bottom quarks $\Delta R_{bb} \geq 0.4$. The latter cut is imposed to guarantee sufficient b separation to employ statistically independent b quark tagging. The overall scalar transverse momentum is imposed $H_T^{\text{parton level}} \geq 1$ TeV,

³ In general, there can also be a model-dependent term $\lambda H \bar{t} \gamma_5 T + h.c.$ in the Lagrangian. Here we assume $\lambda = 0$ for simplicity.

consistent with a similar (but stronger) cut at analysis level, cf. Eq. (8).⁴ The cut on $H_T^{\text{parton level}}$ guarantees a reasonably large fraction of events in the signal regions. Note that this parton level cut on H_T makes it difficult to generate events at NLO, because it acts differently on processes with additional jets at matrix element level (the set of partons which contribute to the sum is different). Matching of matrix element with additional jets is also difficult for the same reason. Therefore, we generate background events at LO without matching to higher multiplicities at matrix element level. Thus, the absolute numbers of the background events should be taken with a grain of salt. The generated signal events do not suffer from this approximation.

B. The analysis

We present an analysis that aims to identify the fully hadronic final state $t\bar{t}HH$ from vectorlike top pair production. We do not rely on large-radius “fat” jets and their substructure, which has become a standard approach whenever boosted heavy particles are involved. Conversely, the approach presented here focuses on separately resolved (small-radius) jets and is intended as a proof-of-concept in a realistic and relevant process.

The proposed analysis consists of the following steps, each of which is described in detail in the remainder of this subsection.

1. Event preselection cuts:

Scalar transverse momentum $H_T \geq 1400 \text{ GeV}$ and number of b -tagged jets $\#b \geq 4$.

2. Jet reconstruction and cut $\#\text{jets} \geq 10$. Here, we use several different benchmark algorithms including the mass-jump algorithm.

3. Assignment of jets to the four buckets B_{t1} , B_{t2} , B_{H1} , and B_{H2} and cuts.

4. Kinematic reconstruction of the vectorlike tops, depending on the number of identified top and Higgs buckets.

⁴ To determine the respective cross-section at large scalar transverse momentum, we cut on events generated with $H_T^{\text{parton level}} \geq 500 \text{ GeV}$ to achieve better accuracy. Only for plotting we also generate $t\bar{t}$ and $b\bar{b}b\bar{b}$ events with $H_T^{\text{parton level}} \geq 1.2 \text{ TeV}$.

1. Event preselection

The decay cascade of a heavy vectorlike top pair leads to an energy deposit of $H_T \sim \mathcal{O}(2m_T)$ in the detector. H_T is the scalar transverse momentum, defined as

$$H_T \equiv \sum_{\text{jets } j} p_{\perp}^{(j)}. \quad (7)$$

We require

$$H_T \geq 1400 \text{ GeV} \quad (8)$$

to retain the majority of signal events for a vectorlike top with $m_T \sim 1 \text{ TeV}$ while strongly suppressing all non-resonant background processes. In the event preselection, jets are clustered with the anti- k_T algorithm as implemented in FASTJET 3.0.6 [37] with parameters $R = 0.4$ and $p_{\perp} \geq 20 \text{ GeV}$. Note that the jets are reconstructed differently after the preselection.

As the signal process contains six b quarks in the final state, we also cut on the number of bottom tags. Tagging is performed by DELPHES using the jets defined above. We select a conservative working point where 70% of b -initiated jets are identified correctly, $\epsilon_{\text{tag}} = 0.70$, and assume the mistag rates of charm-initiated jets to be $\epsilon_{\text{mis}}^{(c)} = 0.10$, and $\epsilon_{\text{mis}}^{(udsg)} = 0.01$ for light (quark- or gluon-initiated) jets. Cutting on

$$\#b \geq 4 \quad (9)$$

reduces the relevant backgrounds to b -rich processes with high multiplicity, $pp \rightarrow t\bar{t}$, $pp \rightarrow t\bar{t}b\bar{b}$, $pp \rightarrow b\bar{b}b\bar{b}$, and $pp \rightarrow t\bar{t}H$.

2. Jet reconstruction

Jets are reconstructed from all calorimeter towers that lie within $|\eta| < 4.9$. To avoid “chopped” jets at the boundary of the detector, we require $|\eta_{\text{jet}}| < 4.0$ so that all jets are sufficiently central. The key ingredient to this analysis is the choice of jet clustering algorithm. In our study, we adopt the following benchmark algorithms and compare them:

- A C/A-like mass-jump clustering algorithm with parameters

$$[\text{MJ06}] : (R = 0.6, \quad p_{\perp} \geq 25 \text{ GeV}, \quad \theta = 0.7, \quad \mu = 50 \text{ GeV}). \quad (10)$$

$$[\text{MJ10}] : (R = 1.0, \quad p_{\perp} \geq 25 \text{ GeV}, \quad \theta = 0.7, \quad \mu = 50 \text{ GeV}). \quad (11)$$

- A standard setup with the Cambridge-Aachen algorithm and commonly used clustering parameters,

$$[\text{CA03}] : (R = 0.3, \quad p_{\perp} \geq 25 \text{ GeV}). \quad (12)$$

$$[\text{CA04}] : (R = 0.4, \quad p_{\perp} \geq 25 \text{ GeV}). \quad (13)$$

The minimum jet p_{\perp} is set to be the same to allow for easy comparison of the results. The additional veto parameters specific to mass-jump clustering, θ and μ in Eqs. (10) and (11), are motivated by the results obtained from boosted top quarks [6]. The mass-jump veto leads to jets whose effective radius can vary, and it is this inherent flexibility that will lead to improved results compared to standard jet clustering with fixed angular size.

Because some jets reconstructed with the mass-jump algorithm may experience a very large effective radius,⁵ contamination from pile-up and underlying event can pose problems in a realistic environment. We therefore apply a trimming [38] stage. For each jet j , the constituents are re-clustered with a smaller radius R_{trim} , yielding hard and possibly also soft subjects. The jet is then re-built only from the subjects i that are hard enough,

$$p_{\perp,i} > f_{\text{trim}} p_{\perp,j}. \quad (14)$$

We choose $R_{\text{trim}} = 0.2$ and $f_{\text{trim}} = 0.03$ as suggested in Ref. [38]. Trimming is applied to all the benchmark points, MJ06, MJ10, CA03, and CA04.

After the jets are reconstructed, we require

$$\#\text{jets} \geq 10, \quad (15)$$

three for each top quark and two for each Higgs boson.

3. Bucket construction and tagging

In order to keep the combinatorial choices of this multi-jet process at a manageable level, we make use of the idea of buckets [8, 9]. First of all, the first top bucket B_{t1} is filled with the three jets that minimize

$$\Delta = |m_{\text{bucket}} - m_t|. \quad (16)$$

⁵ This effect will be investigated later, cf. Fig. 4.

Here and for all other buckets, we limit the allowed jet combinations to those which fulfill

$$p_{\perp,\text{bucket}} \geq 200 \text{ GeV}. \quad (17)$$

This prevents wrong assignments from widely separated low-energy jets, which are possible due to the sheer number of possible choices. In addition, only combinations with minimum mutual jet separation

$$\Delta R(j, j) \geq 0.3 \quad (18)$$

are considered, because smaller distances cannot reasonably be resolved by the hadronic calorimeter any more. This cut is also consistent with the cuts applied on generator level, cf. Sec. III A. Note that we do not impose an upper cut on angular spread of the top decay products, as is implicitly done in all substructure methods which rely on a fat jet of fixed radius. Also note that Eq. (18) does not restrict the analysis if all jets are mutually separated by more than $\Delta R(j, j) = 0.3$, i.e. the fixed- R setups CA03 and CA04 are unaffected. If two top subjets are very close-by and merge in the CA03 setup, even in the ideal case that the MJ algorithm can resolve them separately, they could not contribute to the same bucket. In this sense the cut helps to allow a fair comparison between the mass-jump setups and the Cambridge-Aachen setups.

After the first top bucket has been fixed, out of the remaining jets the second top bucket B_{t2} is filled with three jets, then the first Higgs bucket B_{H1} with two jets, and finally B_{H2} again with two jets. This course of action corresponds to a global metric with explicitly decoupled buckets as defined in Eqs. (4) and (5). Again for each bucket, out of all possible jet combinations that fulfill Eqs. (17) and (18), the combination with minimum metric (Eq. (16), where m_t is replaced by m_H for Higgs buckets) is selected. Events where there is no viable jet combination for a bucket are negligibly rare. Remaining jets are assigned to a fifth bucket B_{ISR} and not further considered in our analysis.

Only after all buckets have been filled, cuts are applied. For top candidate buckets, we require

$$\Delta \leq 25 \text{ GeV}, \quad (19)$$

$$\left(\frac{m_W}{m_t} \right)_{\text{bucket}} \in \frac{m_W}{m_t} \pm 15\%, \quad (20)$$

$$\left(\frac{m_{23}}{m_t} \right)_{\text{bucket}} \geq 0.35. \quad (21)$$

	SR1	SR2	SR3
number of tagged t	= 1	= 2	= 2
number of tagged H	= 2	= 1	= 2

TABLE I: The three signal regions.

Eq. (19) is a simple cut on the reconstructed top mass. The mass ratio in the left-hand side of Eq. (20) is constructed from the two jets which best reconstruct the W boson mass ($m_{W,\text{bucket}}$) and the total jet mass of the bucket ($m_{t,\text{bucket}}$), as proposed when the bucket algorithm was introduced in Ref. [8]. The final cut in Eq. (21) was introduced in the HEPTopTagger [7], where m_{23} is the combined mass of the two sub-leading jets in the bucket (in terms of p_\perp). In our study, it helps to suppress top candidates whose momentum is dominated by one very hard prong.

Higgs candidate buckets have to fulfill

$$\Delta \leq 20 \text{ GeV} . \quad (22)$$

The 4-momentum of the successful top or Higgs candidate is given by the momentum sum of the jets inside the bucket.

4. Signal regions and kinematic reconstruction

We define three signal regions depending on the number of tagged buckets, see Tab. I. In addition to event rates, we kinematically reconstruct the vectorlike top from the momenta of a tagged top quark and a Higgs boson to assess its invariant mass

$$M(t, H) = \sqrt{(p_t + p_H)^2} . \quad (23)$$

In the case of a fully reconstructed event (SR3), we choose between the two possible pairings, $\{(t_1, H_1), (t_2, H_2)\}$ and $\{(t_1, H_2), (t_2, H_1)\}$, such that the mass difference of the two vectorlike tops is minimal,

$$\min [|M(t_1, H_1) - M(t_2, H_2)|, |M(t_1, H_2) - M(t_2, H_1)|] . \quad (24)$$

The majority of events, however, falls into signal regions 1 and 2, and we are left with three tagged and one untagged bucket. As the untagged bucket also contains a significant

energy deposit, its momentum can be used as an estimate for the fourth particle. Again we apply Eq. (24) to determine the correct pairing. Only the vectorlike top that is reconstructed from two tagged buckets is further considered.

C. Results

The cut flow and expected event numbers at the LHC14 with 100 fb^{-1} integrated luminosity are shown in Tab. II for two benchmark setups MJ10 [Eq. (11)] and CA03 [Eq. (12)]. For both setups, the $T\bar{T}$ signal outnumbers the SM background up to $m_T = 800 - 900 \text{ GeV}$ in signal region the SR2, and up to $m_T = 1.1 \text{ TeV}$ in the signal regions SR1 and SR3. Note that the absolute numbers should be taken with great care due to the simplified event generation setup, cf. Sec. III A. A comparison of the relative significances between the employed algorithms is less affected by the uncertainties, though.

We observe that event numbers are largest in SR2 (2 tagged top quarks, 1 tagged Higgs boson). This is particularly pronounced for the various background processes. It can be understood by the order in which the four buckets are filled: The top buckets are filled first and reconstruct the truth partons very well, as will be investigated in Sec. IV. If jets originating from a Higgs boson are wrongly assigned to a top bucket, it becomes unlikely to fill both Higgs buckets from the remaining jets with masses within the mass window. This effect is larger for background processes, among which only a vanishing fraction contains actual Higgs bosons at parton level except the $t\bar{t}H$ background. Thus the Higgs buckets are dominantly filled from the remaining unrelated jets.

As can be seen in Tab. II, in the conventional clustering setup CA03, event numbers are considerably smaller than those obtained with mass-jump clustering MJ10, both for the signal and SM backgrounds. This is already observable at the $\#\text{jets} \geq 10$ cut stage, and the difference becomes even larger when events in the final signal regions are compared. Due to the fixed jet radius of CA03, hard prongs that are separated by a distance smaller than $R = 0.3$ merge, and it is easily understood that the number of hard jets is naturally smaller than the one obtained from a (reasonable) mass-jump setup. As our implementation of the bucket algorithm explicitly requires resolved constituent jets, those merged jets fail to reconstruct their hard resonance, leading to a large drop in event numbers in all signal regions.

Process	$T\bar{T}$					b.g.	$bb\bar{b}\bar{b}$	$t\bar{t}$	$t\bar{t}b\bar{b}$	$t\bar{t}H$
	800 GeV	900 GeV	1.0 TeV	1.1 TeV	1.2 TeV					

number of events for 100 fb^{-1}

$H_T \geq 1.4 \text{ TeV}$		507	306	167	86.7	43.6	25600	4130	20600	772	52.9
$\#b \geq 4$		356	217	118	60.8	30.6	1730	990	506	218	16.4
MJ10	$\#\text{jets} \geq 10$	306	185	101	52.7	26.8	518	166	201	141	9.5
	SR1	14.9	10.4	6.4	3.8	1.9	3.3	0.5	1.1	1.5	0.1
	SR2	36.5	20.7	11.5	5.7	2.7	22.2	1.8	11.4	8.1	0.9
	SR3	10.5	6.0	3.9	2.2	1.0	2.1	0.1	1.1	0.8	0.1

CA03	$\#\text{jets} \geq 10$	282	172	90.6	45.6	22.9	392	121	145	118	7.7
	SR1	8.4	8.0	4.7	2.4	1.2	2.4	0.4	1.0	0.9	0.1
	SR2	24.4	14.1	6.8	3.6	1.8	11.1	0.8	5.3	4.5	0.5
	SR3	5.6	2.9	1.8	1.1	0.4	0.8	0.0	0.4	0.4	0.1

TABLE II: Expected event numbers for two benchmark setups, mass-jump clustering MJ10 [Eq. (11)] and standard Cambridge-Aachen clustering CA03 [Eq. (12)]. Numbers are given for an integrated luminosity of 100 fb^{-1} at the LHC with $\sqrt{s} = 14 \text{ TeV}$. Results for the signal are shown separately for different values of the vectorlike top mass ranging from 800 GeV to 1.2 TeV. All relevant background processes as well as their sum (“b.g.”) are given in the right-hand columns. The three signal regions (SR) are defined in Table I.

In Fig. 2, we show the distributions of the vectorlike top mass, where stacked histograms of all three signal regions SR1 – SR3 are presented. (In SR3, each event gives two entries.) The kinematic reconstruction of the vectorlike top works very well, as manifest in a clear peak in the figures.

In order to compare the different jet clustering setups, it is instructive to look at signal significance S/\sqrt{B} , which we take from the number of signal events S and number of background events B summed over all three signal regions. Numbers are given in Tab. III for all considered setups. It is observed that among the standard clustering setups CA03 and CA04, the smaller jet radius yields better results. The reason is that nearby prongs can only

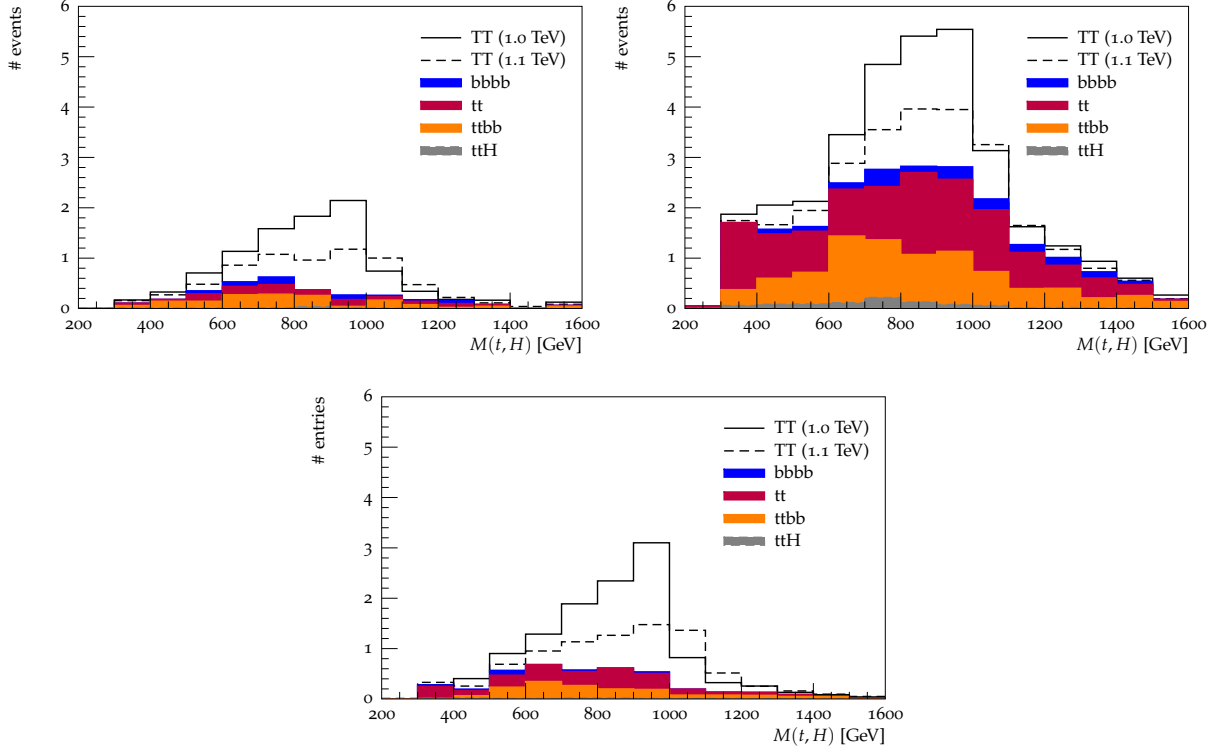


FIG. 2: Reconstructed mass of the vectorlike top $M(t, H)$ (for truth $m_T = 1$ TeV and 1.1 TeV) with the MJ10 setup for SR1 (top left), SR2 (top right) and SR3 (bottom), for an integrated luminosity 100 fb^{-1} . The histograms are stacked.

be separately resolved if the radius parameter is smaller than the mutual separation. For mass-jump clustering MJ10 and MJ06, the opposite is true: the larger maximum jet radius gives more significant results. Even in very busy final states some prongs end up fairly isolated, and they are more accurately reconstructed with larger jets. Overall the mass-jump algorithm outperforms the fixed-radius conventional clustering.

The reconstructed signal mass (for truth $m_T = 1$ TeV) is shown in Fig. 3 for all setups. A peak is visible for all jet clustering setups, but for the fixed-radius algorithms CA03 and CA04 it is shifted to lower values in the SR1 and SR2. The reconstruction is worst for the CA03 setup. Only the analysis based on the mass-jump clustering can reproduce the mass of the heavy T in all signal regions. Independent of the specific clustering algorithm, the reconstructed mass peak has an edge around the true mass, with the majority of events experiencing a lower value. This may be due to the fact that we do not explicitly include

S/\sqrt{B}	800 GeV	900 GeV	1.0 TeV	1.1 TeV	1.2 TeV
MJ10	11.77	7.05	4.13	2.22	1.07
MJ06	11.38	6.96	4.06	2.16	1.02
CA03	10.17	6.63	3.49	1.86	0.90
CA04	11.06	5.91	3.36	1.51	0.61

TABLE III: Comparison of significance S/\sqrt{B} (number of signal events S and number of background events B summed over all three signal regions) for different jet algorithms and benchmark setups.

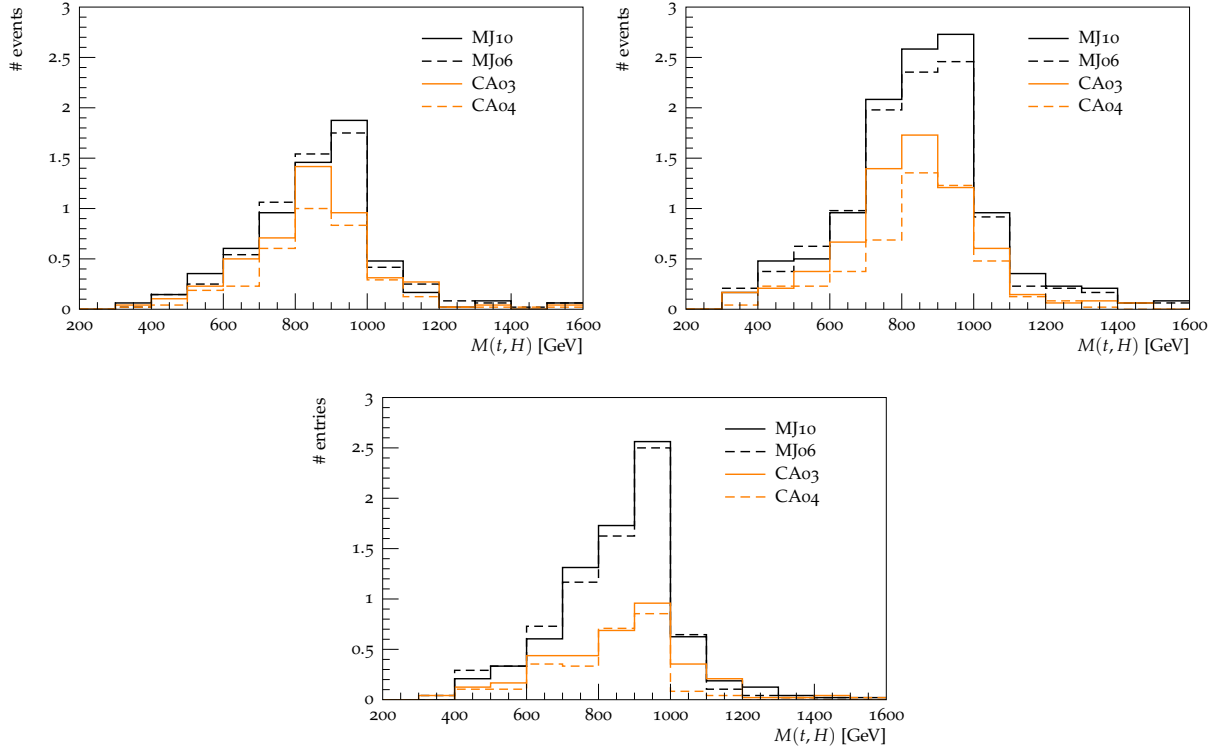


FIG. 3: Reconstructed T mass (for truth $m_T = 1$ TeV) in SR1 (top left), SR2 (top right) and SR3 (bottom) for different jet algorithms, for an integrated luminosity 100 fb^{-1} .

the leading gluon emission when the buckets are reconstructed.

Possible explanations for these observations and a comparison between standard Cambridge-Aachen and the mass-jump jets are given in the following subsection.

D. Comparison of jet clustering algorithms

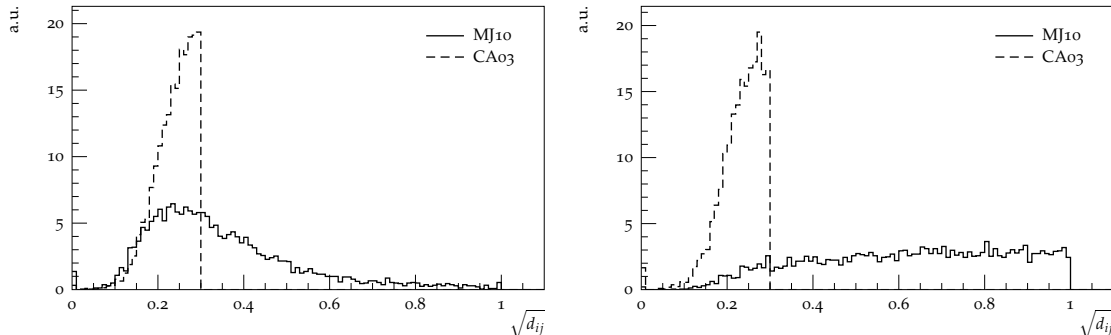


FIG. 4: $\Delta R = \sqrt{d_{ij}}$ of the last recombination step in the hardest (left) and tenth-hardest jet (right) in signal events with $m_T = 1$ TeV (arbitrary units). The solid lines depict values for jets clustered with the C/A-like mass-jump algorithm (MJ10), whereas jets clustered with the conventional C/A algorithm (CA03) are given by dashed lines.

The results found in the previous subsection have mixed implications for the ideal jet radius when standard fixed- R clustering is employed. CA03 yields larger overall significance than CA04, cf. Tab. III. This is not surprising, as only a small radius can separately resolve hard prongs from boosted top and Higgs decays. In terms of event numbers, this advantage seems to well compensate for possible splash-out, a loss of final-state radiation that falls outside the cone. On the other hand, Fig. 3 shows that the reconstruction of the vectorlike top mass works better with a larger radius. This shows the difficulty to find an optimal radius R in the fixed- R clustering algorithms.

Instead of employing a fixed clustering radius, the mass-jump algorithm was designed to separately resolve hard prongs at any distance scale if the terminating veto is called. Fig. 4 shows the angular distance $\Delta R = \sqrt{d_{ij}}$ of the last recombination step in the hardest (left) and tenth-hardest jet (right). Whereas for CA03 jets (dashed lines) the ΔR distribution peaks at the radius cut $R = 0.3$ or slightly below, MJ10 jets (solid lines) observe much more variety. For the hardest jet, ΔR has a peak at around 0.25 but can also take a large value, and for the tenth-hardest jet it has a broader and almost flat distribution. This inherent flexibility constitutes the key to reconstructing the busy final state considered here.

The tail up to very large values in ΔR seen for the mass-jump tenth (soft) jets (Fig. 4

r.h.s.) may be a relic of the algorithm. In terms of significance, it was nevertheless observed that the overall performance is improved when such large radii are allowed.⁶ These large-area jets can gather additional soft radiation, e.g. soft gluon emissions, which can lead to a more accurate bucket mass. The fixed- R setups CA03 and CA04 do not have these features, which could explain why the reconstructed T mass in SR1 and SR2 is shifted to lower values for these algorithms. We speculate that a dedicated study of this effect may lead to improved taggers in this context, but it is beyond the scope of this paper.

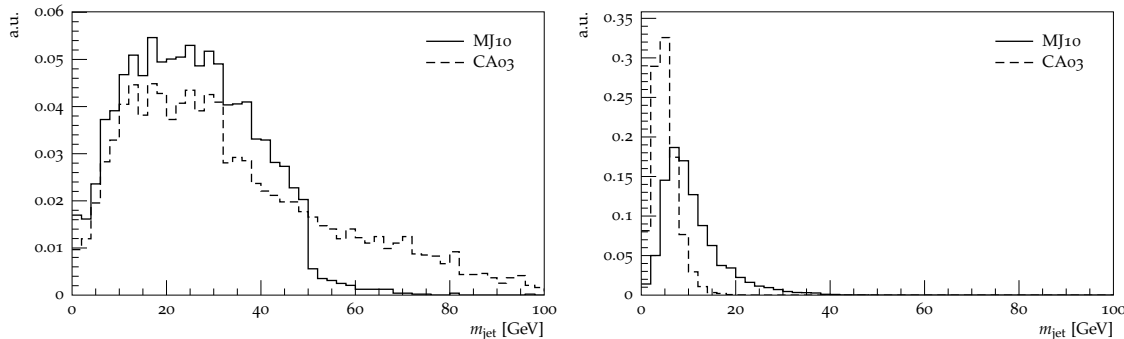


FIG. 5: Trimmed jet mass of the hardest (left) and tenth-hardest jet (right) in signal events with $m_T = 1$ TeV (arbitrary units). The solid lines depict values for jets clustered with the C/A-like mass-jump algorithm (MJ10), whereas jets clustered with the conventional C/A algorithm (CA03) are given by dashed lines.

Fig. 5 shows the trimmed jet mass, again for the hardest (left) and tenth-hardest jet (right). A fraction of events experiences a very heavy leading jet around $m_j = 70 \sim 80$ GeV in the CA03 setup, indicating that nearby hard prongs have merged. The leading mass-jump jet, on the other hand, has a cutoff at $m_j = \mu = 50$ GeV due to the veto condition (cf. Sec. II A), and very large jet masses are absent. As plain jet mass roughly scales with $p_\perp \cdot R$, soft jets clustered with fixed- R algorithms tend to be very light, as shown in the right panel of Fig. 5 for the CA03 setup. However, final-state radiation of low- p_\perp jets is less

⁶ This improvement is diminished due to trimming. By including a trimming stage, we assume that our results are not affected much if additional soft radiation from underlying event and pile-up are taken into account. These effects should be included in a realistic study, but pile-up can only be reliably simulated by the experimental collaborations. We assume that our results, in particular the comparison between conventional jet clustering and mass-jump clustering, are still qualitatively valid in our simplified setup.

collimated and ideally caught in jets with larger radius [39]. In the same figure, it is seen that the tenth (soft) MJ10 jets are heavier due to the larger effective jet radius as observed in Fig. 4.

We conclude that the results found in Sec. III C, namely that a small jet radius can be of advantage for conventional fixed- R clustering algorithms, whereas mass-jump clustering benefits from a very large maximum R , can be explained by looking at jet merging scales and mass distributions. For our process including four boosted resonances and a very busy final state, it is essential to find jets with a flexible algorithm. The mass-jump algorithm avoids the problem of searching for a good compromise for the fixed jet radius parameter and leads to physically more appealing jets. Consequently, it generally outperforms its standard fixed- R counterpart, the Cambridge-Aachen algorithm, in the phenomenological analysis.

E. On fat jet contamination

In this subsection, we briefly discuss the expected performance of algorithms using fat jets in the present process, $pp \rightarrow T\bar{T} \rightarrow t\bar{t}HH \rightarrow 10$ jets. In Sec. I, we argued that the fat jets are not well-separated in such a busy hadronic final state, and this problem is illustrated in Fig. 1 for the signal process with vectorlike top mass $m_T = 1$ TeV. The figure shows the angular distributions of the ten partonic (anti)quark final state (Monte Carlo truth) daughters. The black solid (dashed) line shows the distribution of the largest angular distance between the truth daughters of one top quark (Higgs boson) found in an event. The smallest distance between any truth daughters not coming from the same mother resonance is given by the red line. It is observed that the distance between the nearest daughter particles coming from different mother resonances, $\Delta R(\text{NN})$, is typically smaller than the angular spread of a t or H decay, $\Delta R_{\text{max}}(t/H)$. As a result, the fat jets will be contaminated.

To be more specific, we take the default fat jet clustering parameters of the widely used HEPTopTagger [7]

$$\text{Cambridge-Aachen :} \quad R^{\text{fat jet}} = 1.5 \quad \text{and} \quad p_{\perp}^{\text{fat jet}} \geq 200 \text{ GeV} \quad (25)$$

and give some concrete results for the process $pp \rightarrow T\bar{T} \rightarrow t\bar{t}HH \rightarrow 10$ jets ($m_T = 1$ TeV) in Fig. 6 (CA15, upper panels). A different choice of parameters is suggested in comparisons

between boosted top tagging algorithms [1],

$$\text{anti-}k_T : \quad R^{\text{fat jet}} = 1.0 \quad \text{and} \quad p_{\perp}^{\text{fat jet}} \geq 200 \text{ GeV} , \quad (26)$$

and we also give the plots for these fat jets in Fig. 6 (AKT10, lower panel).

The upper panels imply that the clustering radius of the CA15 jets is too large in this situation. The upper left panel shows the distribution of the number of fat jets. In more than 50% of events, only three fat jets are found, and even less in another 20% (Fig. 6 upper left). A fat jet is labelled “pure truth” t (H) if all truth daughter partons of one top quark (or Higgs boson) and no other truth daughter partons are ghost-associated [40].⁷ In each bin, the fraction of pure truth t (H) fat jets are represented by the hatched (black) area. One can see that only a relatively small fraction of the fat jets are pure truth ones. A phenomenological study would have to rely on events with only three fat jets, and the plain jet mass of those leading fat jets is depicted in the upper right panel. There is a large tail towards very large jet masses, which suggests that there is a significant amount of splash-in from jets not coming from the same t or H resonance.

The second setup (AKT10) with a smaller-radius fat jets, shown in the lower row of Fig. 6, behaves better in this respect. In roughly 40% of events the correct number of four fat jets is identified (lower left), although three-jet events are still dominant. For the events with three fat jets, the leading jet mass is shown in the lower left panel. It can be seen that the distribution still shows a tail, but now large jet masses are much less present than in CA15 jets, implying fewer contamination through splash-in. On the other hand, when we compare the fraction of pure jets in events with four fat jets between the two setups, we observe that AKT10 jets behave worse: While almost 75% of the respective CA15 fat jets are pure, this number is degraded to 65% for AKT10.

We conclude that the study of this process is difficult if we rely on fat jets.⁸ The problem of insufficient separation of the boosted resonances (t and H) is not generically avoided even if a different fat jet radius is chosen – the smallest distance between truth daughters from different mothers is typically smaller than the angular spread of the top quark and Higgs

⁷ The truth partons’ momenta are rescaled to infinitesimal p_{\perp} and energy while η and ϕ are kept fixed (“ghosts”), and participate in jet clustering. Those partons that end up as constituents of a certain jet are called ghost-associated. Due to the vanishing energy of the ghosts, the final jets are unaffected.

⁸ For $m_T \lesssim 900$ GeV, an analysis based on fat jets can still reconstruct the vectorlike top [15]. An experimental analysis of the same process also relies on fat jets [16].

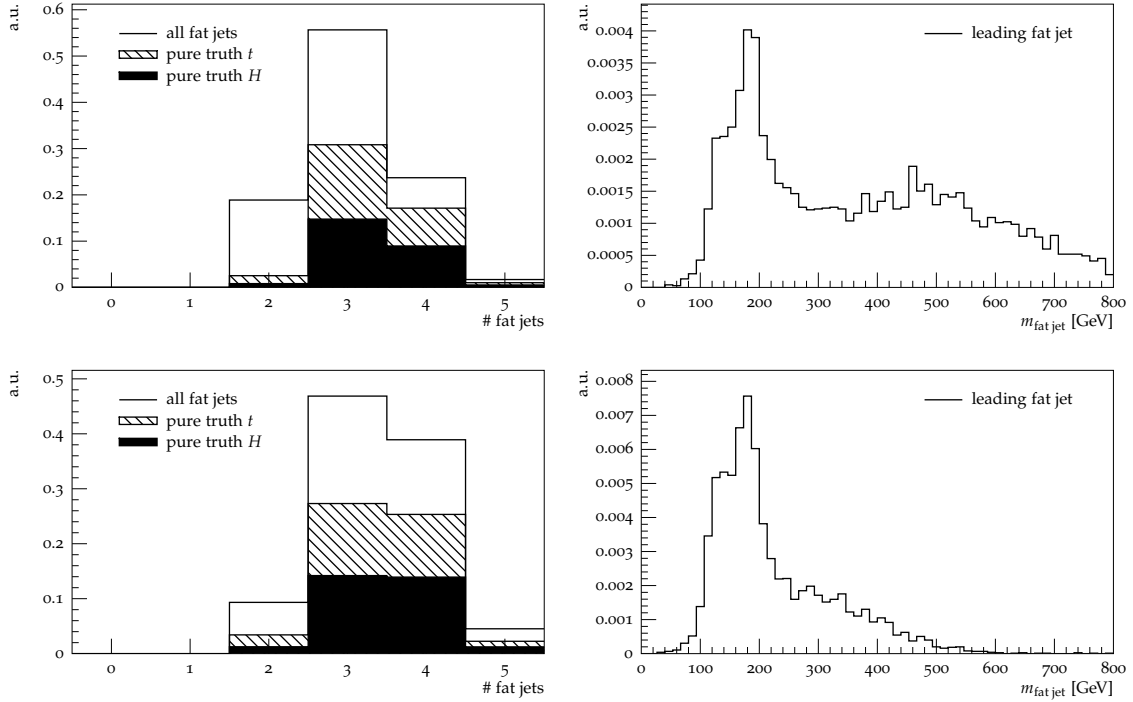


FIG. 6: The upper row shows the results for CA15 fat jets: The number of fat jets is given in the left plot. A fat jet is labelled pure truth t (H) if all truth daughter partons of one top quark (Higgs boson) and no other truth daughter partons are ghost-associated. For events in which three fat jets were found, the distribution of the plain jet mass of the leading fat jet is shown in the right plot. Lower row: The same plots for AKT10 fat jets.

boson, as shown in Fig. 1. There is no apparent solution to this contamination within fat jet algorithms, and the choice of clustering algorithm and parameters is related to finding a balance between splash-in (the fat jet contains energy deposit from a different resonance, too) and splash-out (the fat jet does not contain all radiation from a given resonance). As demonstrated in the main part of this paper, this problem is reduced when the mass-jump algorithm is used.

IV. PERFORMANCE OF TOP AND HIGGS TAGGING

In this section, we investigate the performance of top and Higgs tagging in our approach with the mass-jump (and the Cambridge-Aachen) clustering algorithms. We also briefly

comment on the metric of the decoupled buckets in Sec. IV B.

It should be emphasized that the tagging efficiencies and the quality of reconstruction strongly depend on the considered physical processes as well as the event generator by which the samples have been produced. This is even more true in our study, as the performance of top/Higgs tagging is affected by the hadronic activity from other top and Higgs decay products in the candidate’s vicinity. Tagging efficiencies and reconstruction qualities of the present canonical tagging algorithms for boosted resonances are usually evaluated for isolated fat jets (see e.g. Refs. [1, 2]), which reduces the dependence on the specific process and makes it possible to compare the results between different algorithms. This condition is not satisfied in our benchmark analysis and therefore the results for top and Higgs tagging can hardly be related to other algorithms. In addition, the strong weighting of the global buckets metric in Eq. (4) naturally leads to the first top bucket being much better reconstructed than the second. Similarly, the reconstruction quality of the top quarks is generally better than that of the Higgs bosons.

Despite those precautional warnings, the results presented here can serve as a benchmark for other processes with a similarly busy final state.

A. Reconstruction quality

In Fig. 7 we assess the quality of momentum reconstruction of the tagged buckets for the preferred MJ10 setup, in the benchmark process $pp \rightarrow T\bar{T} \rightarrow t\bar{t}HH \rightarrow 10$ jets at the LHC with $\sqrt{s} = 14$ TeV. The reconstructed masses of the top quarks and Higgs bosons are shown in the upper row. Due to the ordering in the metric, the first bucket always gives a better reconstructed mass, leading to the dip for the second bucket. Both for top and Higgs candidates, there is a central peak at the true mass value for the first buckets. The top mass peak is much narrower, which is not surprising since Higgs buckets are filled by the remaining jets only after the two top buckets have been filled. The middle and lower rows of Fig. 7 show the deviation between the bucket momenta and the MC truth parton momenta in terms of two variables,

$$\frac{\Delta p_{\perp}}{p_{\perp}^{\text{reco}}} \equiv \frac{p_{\perp}^{\text{reco}} - p_{\perp}^{\text{parton}}}{p_{\perp}^{\text{reco}}} \quad \text{and} \quad (27)$$

$$\Delta R_{\text{reco,parton}}. \quad (28)$$

A narrow peak around 0 is observed for both observables, for top as well as Higgs candidates. In fact, most of the tagged buckets are reconstructed within 20% in $\Delta p_{\perp}/p_{\perp}^{\text{reco}}$, and $\Delta R \leq 0.1$, with no significant differences between the respective first and second buckets. We conclude that (i) the tagged buckets are built from the correct jets, and that (ii) these jets reconstruct the truth partons' momenta very well.

For completeness, we also show the same results for the standard clustering benchmark setup CA03 in Fig. 8. Note that, although the distributions look similar to the MJ10 setup, the total number of tagged buckets is significantly smaller. The reconstructed Higgs bosons tend to have a broader peak, shifted to lower values in the CA03 setup. As a result, the mass of the reconstructed vectorlike top is also shifted to lower values, as has been observed in Fig. 3. This is another indication that the mass-jump algorithm is better suited to this analysis, in addition to larger event numbers discussed in Sec. III C.

We conclude this subsection with the reconstruction quality of tagged top buckets from the dominating $t\bar{t}$ SM background (in the MJ10 setup), which is shown in Fig. 9. Deviations from the MC truth partons are very small and our analysis setup is well suited for the background processes as well. We observe that, although transverse momentum is reconstructed generally very accurately, the buckets tend to have lower values than the signal case (cf. Fig. 7). Final-state radiation off the boosted top quark may escape from the respective top bucket, and additional hard prongs from the matrix element that could lead to splash-in are not present.

B. A note on the global metric

The reader might wonder whether the jets are not optimally assigned to the buckets due to the explicitly decoupled metric in Eqs. (4) and (5). This choice was made to reduce the combinatorial workload, but it is expected that the results does not change much even if a more democratic ordered metric, $0 < \omega_{i+1}/\omega_i < 1$, is used. First, we observe that for any bucket the exchange of a jet with one from B_{ISR} does not yield a lower metric by definition, independent of its weights. Secondly, interchange of jets between two buckets (B_i and B_j with $\omega_i > \omega_j$) may lower the measure of B_j , but always at the cost of raising that of B_i . Because of the relative weight $\omega_i > \omega_j$, most of the interchanges are likely to increase the global measure. To find the global minimum, one has to consider a re-assignment of several

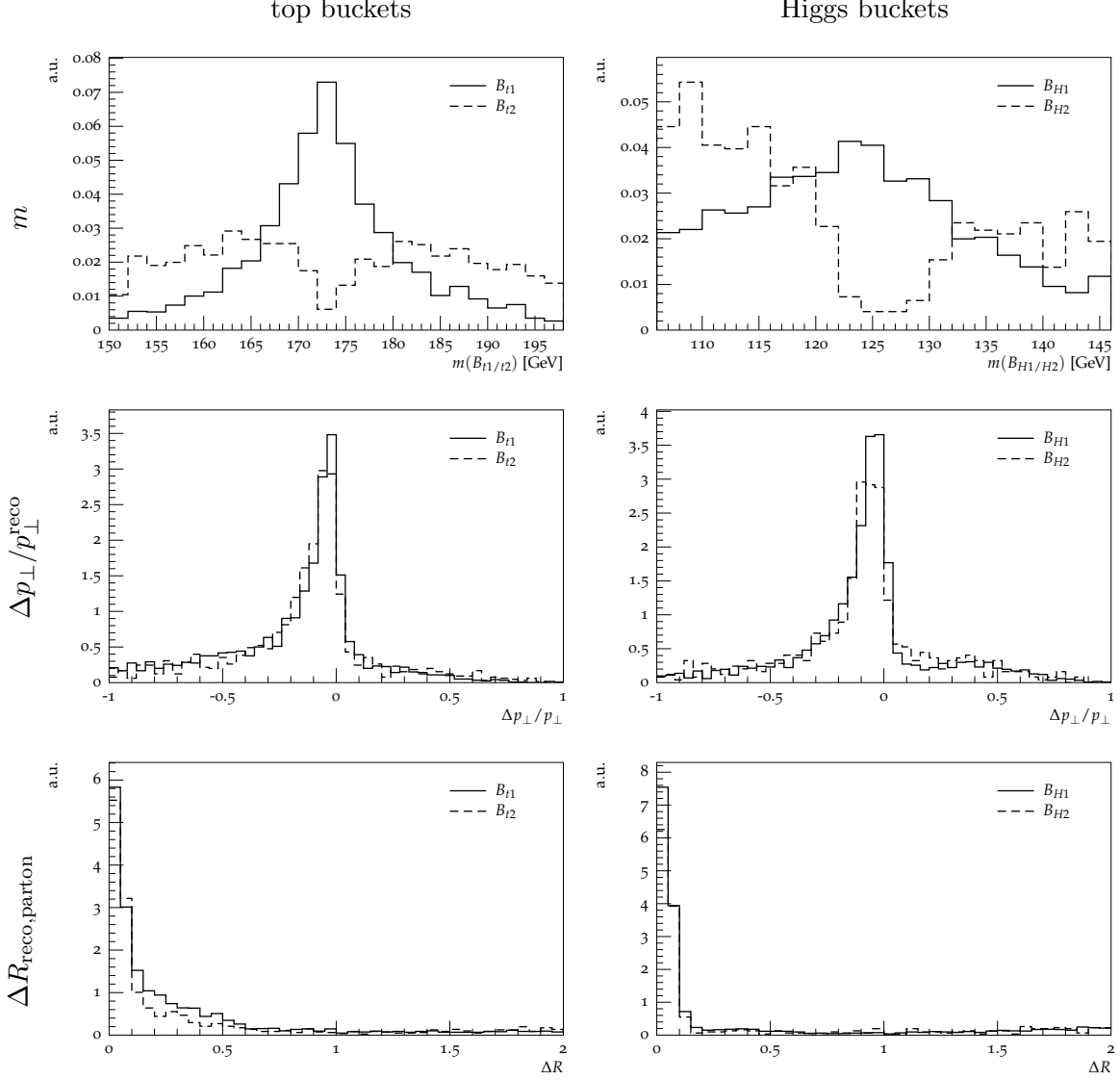


FIG. 7: Reconstruction quality of tagged top (left) and Higgs buckets (right) for the MJ10 setup, in the benchmark process $pp \rightarrow T\bar{T} \rightarrow t\bar{t}HH \rightarrow 10$ jets at the LHC with $\sqrt{s} = 14$ TeV (arbitrary units). From top to bottom, the reconstructed mass, relative deviation in transverse momentum $\Delta p_{\perp}/p_{\perp}^{\text{reco}}$, and the angular distance $\Delta R_{\text{reco,parton}}$ are shown. The solid curves show results for the first bucket, the dashed curves for the second bucket.

jets simultaneously, the details of which depend on the specific weights chosen and is beyond the scope of this paper. Note that, even if finite weights are used, the local minimum found with explicitly decoupled buckets gives an upper bound on Δ_{min}^2 , thus helping to reduce the huge number of permutations.

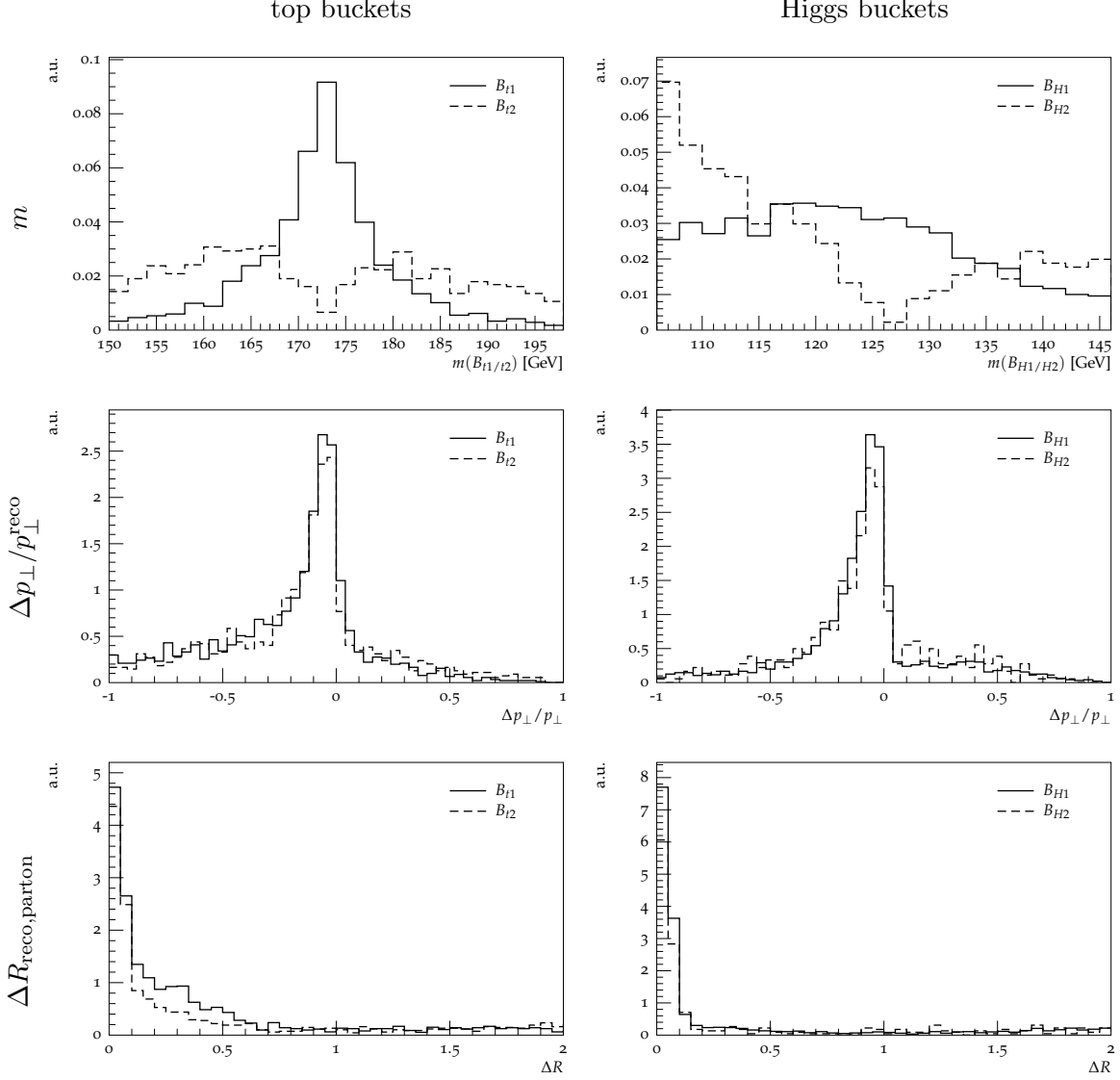


FIG. 8: The same as Fig. 7 but for the CA03 setup.

This argument is weakened if a metric is chosen that does not favour top buckets over Higgs buckets, i.e. $\omega_1, \omega_2 \not\propto \omega_3, \omega_4$. In our analysis, we chose to reconstruct top quarks from three prongs first, and only after that Higgs bosons from two prongs each. This order reduces wrong assignments for both the signal and background. Since SM processes containing Higgs bosons in the final state are rare, the mass distributions in the Higgs buckets can serve as side bands to experimentally determine the background cross-sections.

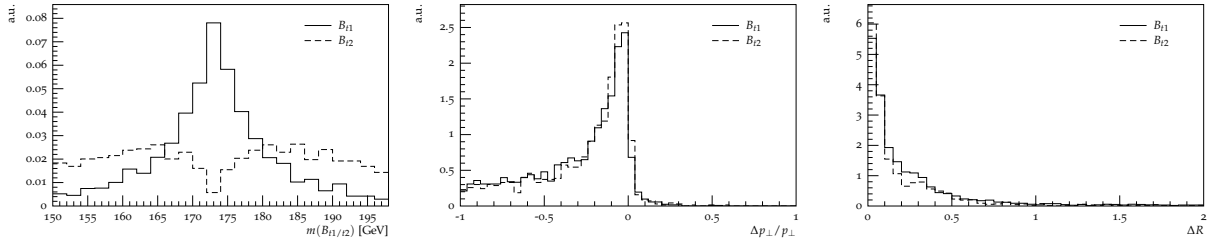


FIG. 9: Reconstruction quality of tagged top buckets of the leading $t\bar{t}$ background events in the MJ10 setup.

V. SUMMARY AND OUTLOOK

We have presented a novel approach to the very busy all-hadronic final state emerging from multiple heavy resonances, focusing on vectorlike top pair production at the LHC as the benchmark of our studies. Since the standard techniques using large-radius fat jets suffer from splash-in contamination and jet overlap in such a busy environment, in this paper we completely relied on separately resolved jets. It was shown that this approach – in combination with a bucket algorithm to reduce computational weight – gives good results and can serve as an alternative channel in new physics searches, including a kinematic reconstruction of the vectorlike top mass. The key ingredient is the mass-jump jet clustering algorithm, which is shown to greatly improve the performance compared to common jet algorithms. This algorithm, which established the family of jet clustering with a terminating veto, is able to resolve nearby hard partons into separate jets, while it resembles common jet algorithms if the partons are well-isolated. In addition to intrinsic jet properties, it introduces a dependence of the clustering history on two-jet properties, all formulated in terms of jet mass and mass ratios. It is this flexibility that outputs jets with variable effective radii, which leads to superior results compared to the fixed-radius variants.

While a χ^2 -like measure could give a more accurate assignment of the jets to the various buckets, we gave an argument that the difference to our computationally inexpensive ansatz is not expected to be large. Another possible improvement is to require a certain number of b -tagged jets for each top and Higgs candidate. We did not include this option in our analysis because bottom tagging is difficult in such busy final states, and also it would require matching between tagged jets and mass-jump jets, which have not yet been investigated by

the experimental collaborations. Our results give a conservative estimate in this respect.

On top of the phenomenological study of vectorlike top pair production, we investigated the quality of reconstruction of the top quarks and Higgs bosons. Whereas the majority of tagging algorithms for boosted resonances assumes their isolation, we showed that our approach performs excellently in identifying the correct jet combinations even in this very busy and unclean environment. This study enters uncharted and often neglected territory when it comes to taggers, yet the results are promising and we expect that jet clustering algorithms with a terminating veto will find their place in future studies of high-multiplicity processes. The algorithm dubbed `ClusteringVetoPlugin` is publicly available in the FASTJET contributions package [29].

Acknowledgments

M.S. is thankful to Michihisa Takeuchi for helpful discussion and advice. We are also grateful to Gavin Salam of the FASTJET team for useful comments on and support with the FASTJET plugin. The work of S.P.L. and M.S. was supported in part by the Program for Leading Graduate Schools, MEXT, Japan. The work of S.P.L. was also supported in part by a JSPS Research Fellowships for Young Scientists. This work was supported by Grant-in-Aid for Scientific research No. 26104001, 26104009, 26247038, 26800123, and also by World Premier International Research Center Initiative (WPI Initiative), MEXT, Japan.

-
- [1] A. Altheimer, S. Arora, L. Asquith, G. Brooijmans, J. Butterworth, M. Campanelli, B. Chapleau and A. E. Cholakian *et al.*, J. Phys. G **39**, 063001 (2012) [arXiv:1201.0008 [hep-ph]].
 - [2] D. Adams, A. Arce, L. Asquith, M. Backovic, T. Barillari, P. Berta, D. Bertolini and A. Buckley *et al.*, arXiv:1504.00679 [hep-ph].
 - [3] M. Dasgupta, L. Magnea and G. P. Salam, JHEP **0802**, 055 (2008) [arXiv:0712.3014 [hep-ph]].
 - [4] M. Cacciari, J. Rojo, G. P. Salam and G. Soyez, JHEP **0812**, 032 (2008) [arXiv:0810.1304 [hep-ph]].

- [5] G. Soyez, JHEP **1007**, 075 (2010) [arXiv:1006.3634 [hep-ph]].
- [6] M. Stoll, JHEP **1504**, 111 (2015) [arXiv:1410.4637 [hep-ph]].
- [7] T. Plehn, M. Spannowsky, M. Takeuchi and D. Zerwas, JHEP **1010**, 078 (2010) [arXiv:1006.2833 [hep-ph]].
- [8] M. R. Buckley, T. Plehn and M. Takeuchi, JHEP **1308**, 086 (2013) [arXiv:1302.6238 [hep-ph]].
- [9] M. R. Buckley, T. Plehn, T. Schell and M. Takeuchi, JHEP **1402**, 130 (2014) [arXiv:1310.6034 [hep-ph]].
- [10] R. Barbier, C. Berat, M. Besancon, M. Chemtob, A. Deandrea, E. Dudas, P. Fayet and S. Lavignac *et al.*, Phys. Rept. **420**, 1 (2005) [hep-ph/0406039].
- [11] J. Fan, M. Reece and J. T. Ruderman, JHEP **1111**, 012 (2011) [arXiv:1105.5135 [hep-ph]].
- [12] S. Gopalakrishna, T. Mandal, S. Mitra and G. Moreau, JHEP **1408**, 079 (2014) [arXiv:1306.2656 [hep-ph]].
- [13] A. Girdhar, B. Mukhopadhyaya and M. Patra, Phys. Rev. D **91**, no. 5, 055015 (2015) [arXiv:1404.3374 [hep-ph]].
- [14] S. Yang, J. Jiang, Q. S. Yan and X. Zhao, JHEP **1409**, 035 (2014) [arXiv:1405.2514 [hep-ph]].
- [15] M. Endo, K. Hamaguchi, K. Ishikawa and M. Stoll, Phys. Rev. D **90**, no. 5, 055027 (2014) [arXiv:1405.2677 [hep-ph]].
- [16] V. Khachatryan *et al.* [CMS Collaboration], arXiv:1503.01952 [hep-ex].
- [17] G. Aad *et al.* [ATLAS Collaboration], JHEP **1411**, 104 (2014) [arXiv:1409.5500 [hep-ex]].
- [18] G. Aad *et al.* [ATLAS Collaboration], arXiv:1503.05425 [hep-ex].
- [19] The ATLAS collaboration, ATLAS-CONF-2015-012, ATLAS-COM-CONF-2015-012.
- [20] S. Chatrchyan *et al.* [CMS Collaboration], Phys. Rev. Lett. **107**, 271802 (2011) [arXiv:1109.4985 [hep-ex]].
- [21] S. Chatrchyan *et al.* [CMS Collaboration], Phys. Lett. B **729**, 149 (2014) [arXiv:1311.7667 [hep-ex]].
- [22] CMS Collaboration [CMS Collaboration], CMS-PAS-B2G-12-017.
- [23] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock and B. R. Webber, Phys. Lett. B **269**, 432 (1991).
- [24] S. Catani, Y. L. Dokshitzer, M. H. Seymour and B. R. Webber, Nucl. Phys. B **406**, 187

- (1993).
- [25] S. D. Ellis and D. E. Soper, Phys. Rev. D **48**, 3160 (1993) [hep-ph/9305266].
 - [26] Y. L. Dokshitzer, G. D. Leder, S. Moretti and B. R. Webber, JHEP **9708**, 001 (1997) [hep-ph/9707323].
 - [27] M. Wobisch and T. Wengler, In *Hamburg 1998/1999, Monte Carlo generators for HERA physics* 270-279 [hep-ph/9907280].
 - [28] M. Cacciari, G. P. Salam and G. Soyez, JHEP **0804**, 063 (2008) [arXiv:0802.1189 [hep-ph]].
 - [29] <http://fastjet.hepforge.org/contrib/>
 - [30] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer and T. Stelzer, JHEP **1106**, 128 (2011) [arXiv:1106.0522 [hep-ph]].
 - [31] T. Sjostrand, S. Mrenna and P. Z. Skands, JHEP **0605**, 026 (2006) [hep-ph/0603175].
 - [32] J. de Favereau *et al.* [DELPHES 3 Collaboration], JHEP **1402**, 057 (2014) [arXiv:1307.6346 [hep-ex]].
 - [33] N. Kidonakis and R. Vogt, Phys. Rev. D **78**, 074005 (2008) [arXiv:0805.3844 [hep-ph]].
 - [34] G. Bevilacqua, M. Czakon, C. G. Papadopoulos, R. Pittau and M. Worek, JHEP **0909**, 109 (2009) [arXiv:0907.4723 [hep-ph]].
 - [35] R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, R. Pittau and P. Torrielli, Phys. Lett. B **701**, 427 (2011) [arXiv:1104.5613 [hep-ph]].
 - [36] M. Worek, PoS RADCOR **2013**, 038 (2013) [arXiv:1311.2396 [hep-ph]].
 - [37] M. Cacciari, G. P. Salam and G. Soyez, Eur. Phys. J. C **72**, 1896 (2012) [arXiv:1111.6097 [hep-ph]].
 - [38] D. Krohn, J. Thaler and L. T. Wang, JHEP **1002**, 084 (2010) [arXiv:0912.1342 [hep-ph]].
 - [39] D. Krohn, J. Thaler and L. T. Wang, JHEP **0906**, 059 (2009) [arXiv:0903.0392 [hep-ph]].
 - [40] M. Cacciari, G. P. Salam and G. Soyez, JHEP **0804**, 005 (2008) [arXiv:0802.1188 [hep-ph]].